*Nucleic Acids Research, 2012, Vol. 40, Database issue* **D115–D122**
*doi:10.1093/nar/gkr1044*

# IMG: the integrated microbial genomes database and comparative analysis system

**Victor M. Markowitz[1,*], I-Min A. Chen[1], Krishna Palaniappan[1], Ken Chu[1], Ernest Szeto[1], Yuri Grechkin[1], Anna Ratner[1], Biju Jacob[1], Jinghua Huang[1], Peter Williams[2], Marcel Huntemann[2], Iain Anderson[2], Konstantinos Mavromatis[2], Natalia N. Ivanova[2] and Nikos C. Kyrpides[2,*]**

[1]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley and [2]Microbial Genomics and Metagenomics Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

Received September 15, 2011; Accepted October 24, 2011



## ABSTRACT

**The Integrated Microbial Genomes (IMG) system serves as a community resource for comparative analysis of publicly available genomes in a comprehensive integrated context. IMG integrates publicly available draft and complete genomes from all three domains of life with a large number of plasmids and viruses. IMG provides tools and viewers for analyzing and reviewing the annotations of genes and genomes in a comparative context. IMG's data content and analytical capabilities have been continuously extended through regular updates since its first release in March 2005. IMG is available at http://img.jgi.doe.gov. Companion IMG systems provide support for expert review of genome annotations (IMG/ER: http://img.jgi.doe.gov/er), teaching courses and training in microbial genome analysis (IMG/EDU: http://img.jgi.doe.gov/edu) and analysis of genomes related to the Human Microbiome Project (IMG/HMP: http://www.hmpdacc-resources.org/img_hmp).**

## INTRODUCTION

The Integrated Microbial Genomes (IMG) system integrates publicly available draft and complete microbial genomes from all three domains of life with a large number of plasmids and viruses. IMG employs NCBI's RefSeq resource (1) as its main source of public genome sequence data, and 'primary' annotations consisting of predicted genes and protein products. For every genome, IMG records its primary genome sequence information from RefSeq including its organization into chromosomal replicons (for finished genomes) and scaffolds and/or contigs (for draft genomes), together with predicted protein-coding sequences (CDSs), some RNA-coding genes and protein product names that are provided by the genome sequence centres.

IMG's data integration pipeline associates every genome with metadata from GOLD (2), and fills in additional information potentially missing from the RefSeq files such as CRISPR repeats (3), signal peptides computed using SignalP (4) and transmembrane helices computed using TMHMM (5). Missing RNAs are identified using tRNAS-can-SE-1.23 (6) for tRNAs, in house developed HMMs for rRNAs (7), and Rfam (8) and INFERNAL v1.0 (9) for other small RNAs. Genes are associated with 'secondary' functional annotations and lists of related (e.g. homologue, paralogue) genes. IMG generated annotations consist of protein family and domain characterizations based on COG clusters and functional categories (10), Pfam (11), TIGRfam and TIGR role categories (12), InterPro domains (13), Gene Ontology (GO) terms (14) and KEGG Ortholog (KO) terms and pathways (15).

The association of KEGG pathways with IMG genomes is based on the assignment of KEGG Orthology (KO) terms to IMG genes via a mapping of IMG genes to KEGG genes. The MetaCyc collection of pathways (16) is also available in IMG, whereby the association of MetaCyc pathways with IMG genomes is based on correlating enzyme EC numbers in MetaCyc reactions with EC numbers associated with IMG genes via KO terms. Genes are further characterized using an IMG native collection of generic (protein cluster-independent) functional roles called IMG terms that are defined by their association with generic (organism-independent)

*To whom correspondence should be addressed. Tel: +510 486 7073; Fax: +1 510 486 5812; Email: VMMarkowitz@lbl.gov
Correspondence may also be addressed to Nikos C. Kyrpides. Tel: +925 296 5718; Fax: +925 296 5666; Email: nckyrpides@lbl.gov

functional hierarchies, called IMG pathways (17). IMG terms and pathways are specified by domain experts at DOE-JGI as part of the process of annotating specific genomes of interest, and are subsequently propagated to all the genomes in IMG using a rule based methodology (18). Transporter genes are linked to the Transport Classification Database (19) based on their assignment to COG, Pfam or TIGRfam domains or IMG Terms that correspond to transporter families.

For each gene, IMG provides lists of related (e.g. candidate homologue, paralogue, orthologue) genes that are based on sequence similarities computed using NCBI BLASTp for protein coding genes and BLASTn for RNA genes. Such lists of genes can be filtered using percent identity, bit score and more stringent *E*-values.

IMG's data integration pipeline identifies gene fusions and conserved gene cassettes (putative operons). A fused gene (fusion) is defined as a gene that is formed from the composition (fusion) of two or more previously separate genes (20). Transposases and integrases, pseudogenes, and genes from draft genomes are not considered as putative fusion components in order to avoid false positives caused by gene fragmentation. A 'chromosomal cassette' is defined as a stretch of genes with intergenic distance smaller or equal to 300 bp (21), whereby the genes can be on the same or different strands of the chromosome. Chromosomal cassettes with a minimum size of two genes common in at least two separate genomes are defined as 'conserved chromosomal cassettes'. The identification of common genes across organisms is based on three gene clustering methods, namely participation in COG, Pfam and IMG orthologue clusters (22). Correlation scores between different gene clusters, based on their co-existence on fusion events, conserved chromosomal cassettes and genomes, provide insights in their function (21).

We review below IMG's data content growth and analysis tool extensions since the last published report on IMG (23).

## DATA CONTENT EXTENSIONS

### Genomics data

The content of IMG has grown steadily since the first version released in March 2005, with IMG 3.4 (July 2011) containing 3008 bacterial, archaeal and eukaryotic genomes, an increase of over 80% since August 2009 (23). IMG 3.4 also contains 2697 viral genomes and 1186 plasmids that did not come from a specific microbial genome sequencing project bringing its total genome content to 6891 genomes with over 11.6 million genes (A Content History link on IMG's home page provides an overview of its content growth.).

While archaeal, bacterial, plasmid and viral genomes are updated on a regular basis in IMG, the inclusion of eukaryotic genomes entails a more complex process (The integration process into IMG for eukaryotic genomes is described at: http://img.jgi.doe.gov/w/doc/euks.html.) and is done at longer intervals. Since August 2009, about 70 new eukaryotic genomes have been added to IMG, out of which 40 are fungal genomes.

The 'Expert Review' version of IMG, IMG/ER (24), allows individual scientists or groups of scientists to review and curate the functional annotation of microbial genomes in the context of IMG's public genomes. Scientists can submit their private genome data sets into IMG ER (using password protected access) prior to their public release either with their original annotations or with annotations generated by IMG's annotation pipeline (18). Since August 2009, close to 750 private genomes have been reviewed and curated using IMG/ER.

Genomes generated as part of the Human Microbiome Project (HMP) (25) and the Genome Encyclopedia of Bacterial and Archaea Genomes (GEBA) project (26) are of special interest. With the goal of characterizing microbial communities found at multiple human body sites, HMP has initially focused on the sequencing of reference genomes from both cultured and uncultured bacteria (25). Over 550 reference genomes sequenced as part of the HMP initiative, as well as over 1500 genomes associated with a human host and thus relevant to HMP, can be examined and analyzed using IMG/HMP (http://www.hmpdacc-resources.org/img_hmp/), which is provided as part of the HMP Data Analysis and Coordination Center (DACC).

The aim of the GEBA is to fill systematically the sequencing gaps along the bacterial and archaeal branches of the tree of life. After a pilot project in 2009 that generated complete genomes for about 100 organisms (26), the number of sequenced GEBA genomes has steadily increased and stands at 205 as of August 2011. GEBA genomes are available for analysis or download via a special purpose interface, IMG/GEBA (http://img.jgi.doe.gov/geba/), as soon as their annotation is completed at JGI, and before they are available in Genbank.

### Proteomics data

Proteomics, transcriptomics, metabolomics, epigenomics and interactomics data are increasingly employed jointly with genomics data to refine our understanding of the functions of genes. Accordingly, these types of 'omics' data are gradually included into IMG.

The first protein expression data sets included into IMG were generated as part of the *Arthrobacter chlorophenolicus* study conducted at the Oakridge National Laboratory (27). Subsequently, data sets from Cryptobacterium curtum and Brachybacterium faecium studies conducted at WR Wiley Environmental Molecular Sciences Laboratory, Instrument Development Laboratory, Pacific Northwest National Laboratory were also added to IMG.

For a genome involved in a protein expression study, the experiments/samples are recorded together with the experimental conditions and the protein expression data organized per expressed gene. For each expressed gene, the number of observed peptides is recorded together with peptide sequences and the normalized coverage. The normalized coverage is defined as the coverage of an expressed gene in an experiment divided by the total coverage of the genes in that experiment, where coverage

for a gene is defined as of the number of all observed peptides for the gene divided by the size of the gene (28).

### Predicted phenotypes

Phenotypes are broadly defined as an observable characteristic of an organism. The current list of phenotypes in IMG are predicted using a set of rules based on IMG's native collection of pathways.

Many physiological functions require the coordinated action of several gene products, which can be grouped into pathways, where genes function in a specific order. Pathways can be analyzed in the context of other pathways within the organism. For example, if an organism degrades cellulose to cellobiose outside the cell, it can only utilize cellulose as a carbon source if it also has a transport pathway for uptake of cellobiose and, within the cell, a metabolic pathway to gain energy from cellobiose. If all three steps are present, then the organism has the phenotype of Growth on cellulose via cellobiose. In some cases the presence or absence of only one pathway is required for a phenotype. There are also cases in which there are multiple possibilities and require multiple combinations of pathways.

Phenotype prediction rules consist of AND–OR combinations of IMG pathway assertions. There are currently 56 rules to predict phenotypes grouped into categories and subcategories, as shown in Figure 1 which displays the first 11 rules together with the number of genomes that are associated with a specific phenotype.

## ANALYSIS TOOL EXTENSIONS

Genome data analysis in IMG consists of operations involving genomes, genes and functions which can be selected, explored individually, and compared. The composition of analysis operations is facilitated by genome, scaffold, gene and function 'carts' that handle lists of genomes, scaffolds, genes and functions, respectively.

### Data selection tools

Genomes, genes and functions can be selected using browsers and search tools. Browsers allow users to select genomes and functions organized as alphabetical lists or using domain specific hierarchical classifications. Keyword search tools allow identifying genomes, genes and functions of interest using a variety of selection filters. Genomes can be also selected using a search tool which allows specifying conditions involving metadata attributes, such as temperature range, oxygen requirement or ecosystem, while genes can be also selected using BLAST search tools against various data sets.

IMG's data selection tools have been extended in order to improve their efficiency and usability. For example,

Predicted Phenotypes

| Rule ID | Name | Category | Category Value | Description | No. of Genomes w/ Phenotype |
|---|---|---|---|---|---|
| 00001 | L-histidine prototroph | Metabolism | Prototrophic | Organism is predicted to be able to synthesize L-histidine. | 52 |
| 00002 | Aerobe | Oxygen Requirement | Aerobe | Organism is predicted to be able to grow in the presence of air. | 255 |
| 00003 | L-lysine prototroph | Metabolism | Prototrophic | Organism is predicted to be able to synthesize L-lysine. | 327 |
| 00004 | Denitrifier | Metabolism | Denitrifying | Organism is predicted to be able to reduce nitrate to nitrogen (N2). | 8 |
| 00005 | Use of nitrate as electron acceptor | Metabolism | Nitrate reducer | Organism is predicted to be able to grow anaerobically with nitrate as electron acceptor | 516 |
| 00006 | Carbon fixation | Metabolism | Carbon fixation | Organism is predicted to be able to use carbon dioxide as sole carbon source | 9 |
| 00007 | L-lysine auxotroph | Metabolism | Auxotroph | Organism is predicted to be unable to synthesize L-lysine. | 2322 |
| 00010 | L-alanine prototroph | Metabolism | Prototrophic | Organism is predicted to be able to synthesize L-alanine | 1775 |
| 00011 | L-alanine auxotroph | Metabolism | Auxotroph | Organism is predicted to be unable to synthesize L-alanine | 493 |
| 00012 | L-aspartate prototroph | Metabolism | Prototrophic | Organism is predicted to be able to synthesize L-aspartate | 1701 |
| 00013 | L-aspartate auxotroph | Metabolism | Auxotroph | Organism is predicted to be unable to synthesize L-aspartate | 612 |
| 00014 | L-glutamate prototroph | Metabolism | Prototrophic | Organism is predicted to be able to synthesize L-glutamate | 2344 |
| 00015 | L-glutamate auxotroph | Metabolism | Auxotroph | Organism is predicted to be unable to synthesize L-glutamate | 204 |
| 00016 | L-phenylalanine prototroph | Metabolism | Prototrophic | Organism is predicted to be able to synthesize L-phenylalanine | 231 |

**Figure 1.** A sample of rules for predicting phenotypes in IMG.

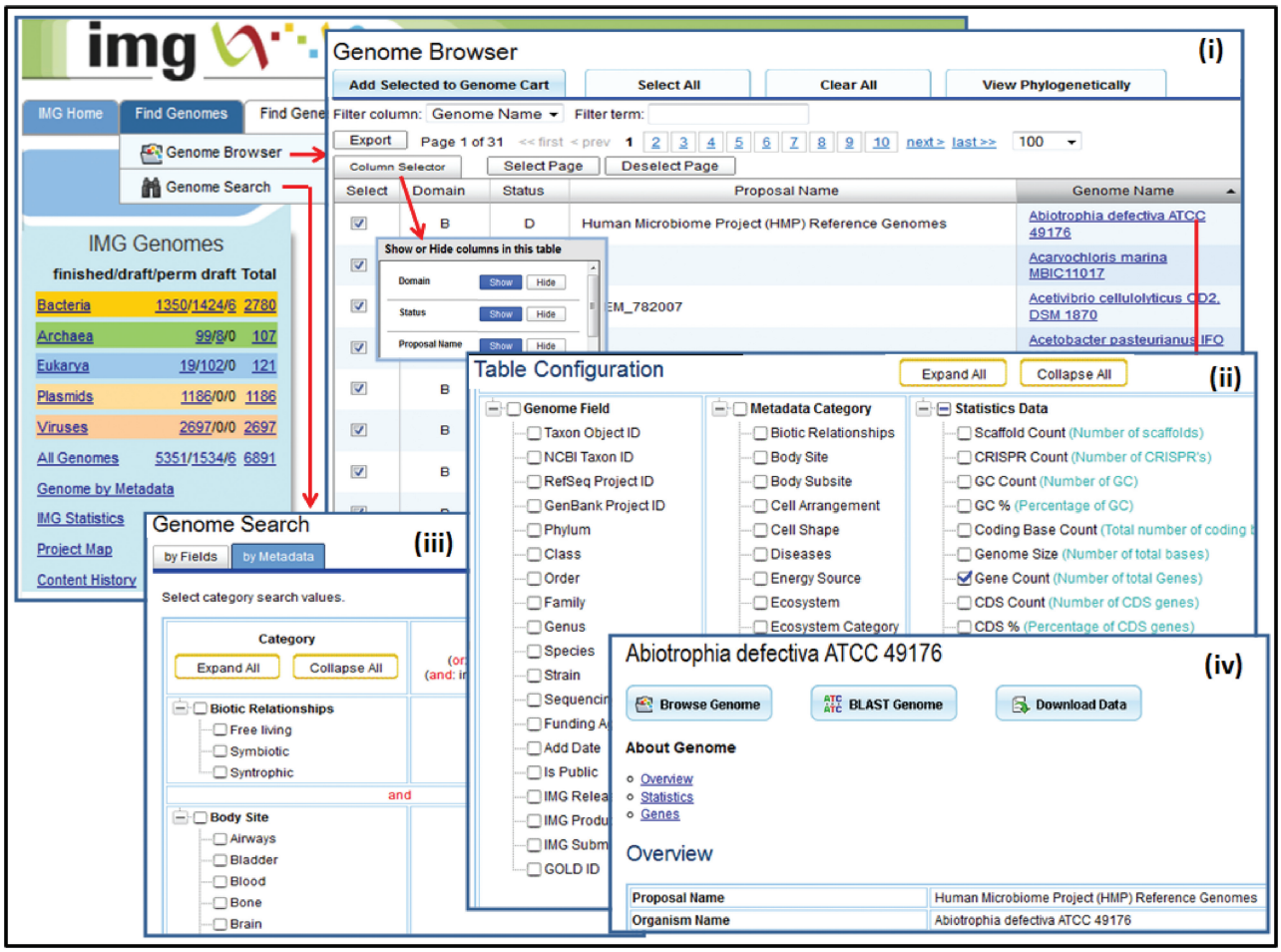


**Figure 2.** Genome browser and search tools. The 'Genome Browser' displays the genomes organized in a phylogenetic tree or (**i**) in a tabular list that can be configured by (**ii**) adding or removing genome, metadata or annotation specific columns. (**iii**) 'Genome Search' allows searching genomes on genome or metadata specific fields. (**iv**) A genome can be explored using a variety of browsing tools, searched for the presence of specific genes using BLAST, or downloaded.

genomes can be selected using 'Genome Browser' or 'Genome Search', as illustrated in Figure 2.

The 'Genome Browser' displays the genomes organized in a phylogenetic tree or in a tabular format as illustrated in Figure 2(i). The tabular display of genomes has a dynamic layout, with columns than can be resized, reordered and sorted on content, configurable page display size, and an export capability for saving tables as Excel spreadsheets or tab delimited files. A 'Column Selector' allows to hide columns. The genome table can be also reconfigured by adding or removing genome, metadata or annotation specific columns, as illustrated in Figure 2(ii). Note that the number of metadata attributes associated with genomes has increased substantially in the past few years, whereby the data for these attributes is collected from GOLD (2). 'Genome Search' allows searching genomes on genome or metadata specific fields, as illustrated in Figure 2(iii).

Individual genomes can be explored using the 'Organism Details' page which provides a variety of tools for browsing, searching for the presence of specific genes, or downloading genome data sets, as illustrated in Figure 2(iv). This page also provides information (metadata) on the genome together with various genome statistics of interest, such as the number of genes that are associated with KEGG, COG, Pfam, InterPro or enzyme information. Individual genes can be analyzed using the 'Gene Details' page which includes Gene Information, Protein Information and Pathway Information tables, evidence for functional prediction, COG, Pfam and pre-computed homologues.

Tabular and graphical displays, such as graphical viewers for the distribution of genes associated with COG, Pfam, TIGRfam and KEGG for each genome, have been extended in order to facilitate genome and gene exploration. Individual functional categories, such as COG, Pfam, TIGRfam, KEGG Orthology terms and pathways, can be explored using functional category specific browsers.

New IMG tools provide support for examining protein expression data as illustrated in Figure 3. Protein expression studies are listed on the 'Experiments Statistics' section of the 'IMG Statistics' page and are available on the 'Organism Details' page of the genome they are associated with. A protein expression study, such as 'Impact of Phenolic Substrate and Growth Temperature

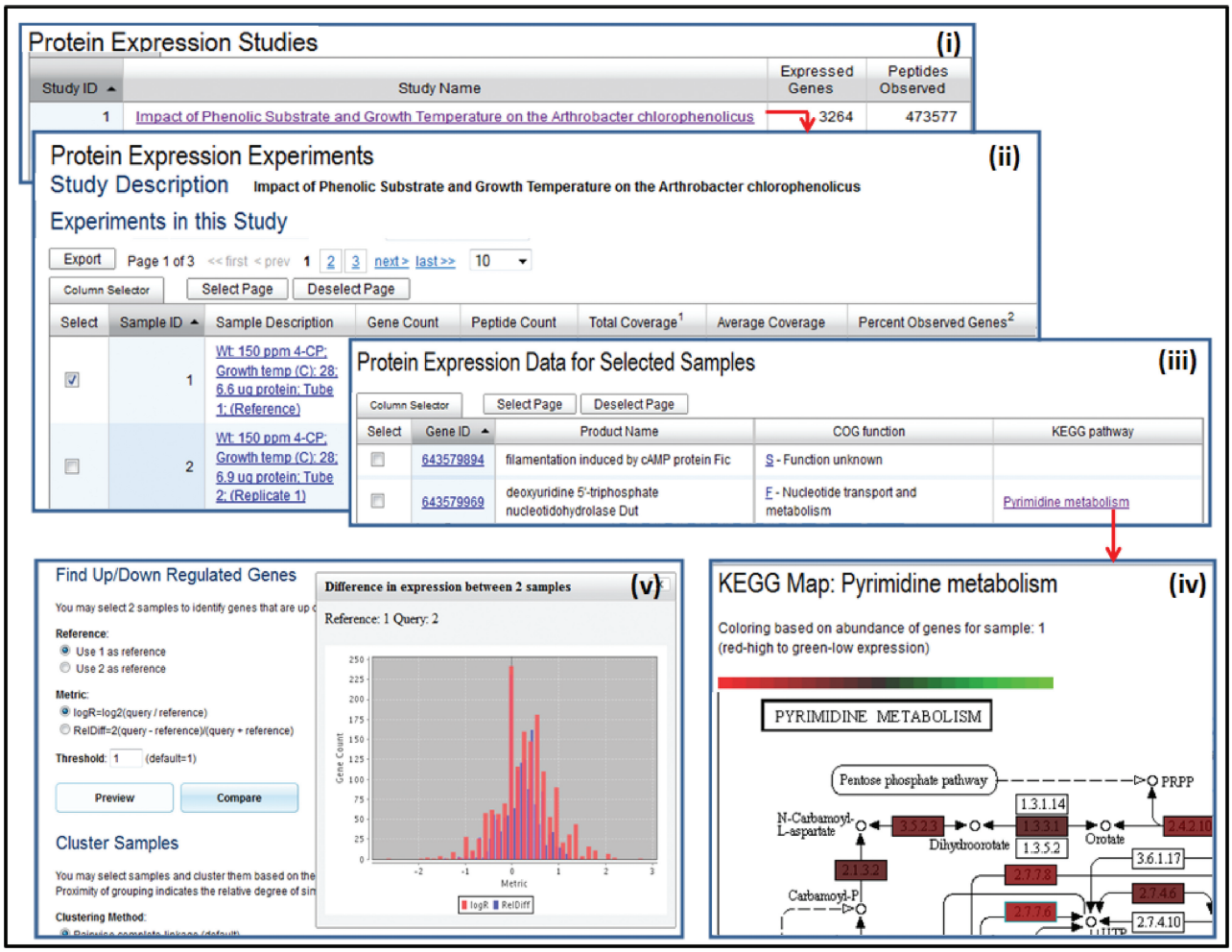


**Figure 3.** Protein expression exploration tools. **(i)** 'Protein Expression Studies' are listed on the IMG Statistics page, with each study associated with **(ii)** a list of 'Protein Expression Experiments' (samples). **(iii)** Samples can be selected for further analysis, such as examining expressed genes of **(iv)** a single sample in the context of pathway, where enzymes are displayed with colours representing the level of expression for the associated genes. **(v)** Sample pairs can be compared in terms of genes up or down regulation, with the result of the comparison displayed as a histogram.

on the Arthrobacter chlorophenolicus' study shown in Figure 3(i), is associated with a list of samples (experiments). Summaries for samples include a description, the number of associated genes, the peptide count and the total and average coverage for the sample (The total coverage is the sum of coverages for the genes in a sample, where the coverage for a gene consists of the count of its associated peptides divided by the size of the gene.), as illustrated in Figure 3(ii). Samples can be selected for further analysis. Expressed genes of a single sample can be examined in the context of pathways, as illustrated in Figure 3(iv), whereby enzymes are displayed with colours representing the level of expression for the associated genes. Expressed genes of multiple samples can be also examined in the context of pathways, whereby enzymes are displayed with colours representing the percentage of samples with expressed genes associated with the enzymes. Samples (experiments) can be clustered based on coverage values for the genes expressed in each sample, with a choice of clustering methods, such as pairwise complete linkage and centroid linkage, and distance measure, such as Pearson correlation, Spearman's rank correlation and Euclidean distance. The result of clustering is displayed as a hierarchical tree of samples and a normalized heat map of coverage values for each gene for each sample.

Sample pairs can be compared in terms of genes up or down regulation, with a threshold specified for the difference in gene expression. The difference in expression is computed using either the $\log R = \log2(\text{query}/\text{reference})$ or the $\text{RelDiff} = 2(\text{query} - \text{reference})/(\text{query} + \text{reference})$ metric. The result of the comparison can be displayed as a histogram, as illustrated in Figure 3(v), or in a tabular format. This histogram can be used to identify and set thresholds for the search of over expressed or under expressed genes between any pair of selected conditions.

The genomes, genes and functions that result from search operations are displayed as lists from which genomes, genes and functions can be selected for inclusion into the 'Genome Cart', 'Gene Cart' and 'Function Cart', respectively. These carts have been extended in order to facilitate the composition of analysis tools in IMG. Thus, genes selected in 'Gene Cart' can be added directly to 'Function Cart' via their associated functions,

such as COG, Pfam, TIGRfam. In a similar manner, functions selected in 'Function Cart' can be added directly to 'Gene Cart' via the genes associated with the selected functions, where the genes included into the 'Gene Cart' can be restricted to specific genomes.

### Comparative analysis tools

Genomes can be compared in terms of gene content using the 'Phylogenetic Profiler' and 'Phylogenetic Profiler for Gene Cassettes' tools. The 'Phylogenetic Profiler' allows users to identify genes in a query genome in terms of presence or absence of homologues in other genomes. The 'Phylogenetic Profiler for Gene Cassettes' allows users to find genes that are part of a gene cassette in a query genome as well as part of related (conserved part of) gene cassettes in other genomes, whereby the result of such a search includes groups of collocated genes in each chromosomal cassette in the query genome that satisfy the search condition. More details on context analysis based on IMG's gene cassettes can be found in (22).

Genomes can be compared in terms of functional capabilities using the 'Abundance Profile Overview' and 'Function Profile' tools. The 'Abundance Profile Overview' allows users to compare the relative abundance of protein families (COGs, Pfams, TIGRfams) and functional families (enzymes) across selected genomes, whereby the results are displayed either as a heat map or a matrix, with the cells in the heat map and matrix linked to the list of genes assigned to a particular family in a genome. The 'Function Profile' is a selective version of the 'Abundance Profile Overview', with functions of interest first selected with the 'Function Cart'.

The metabolic capabilities of genomes can be compared using the 'Abundance Profile Overview' and 'Function Profile' tools applied on enzymes involved in a pathway of interest. Alternatively, the metabolic capabilities of genomes can be compared in the context of KEGG pathways, as illustrated in Figure 4. Once a pathway is selected from the list of KEGG pathways via the KEGG option of the 'Find Functions' menu, as shown in Figure 4(i), the 'KEGG Pathway Details' lists the associated enzymes of KO terms, as illustrated in Figure 4(ii). Genomes for comparison are selected from a phylogenetically organized list, with the comparison

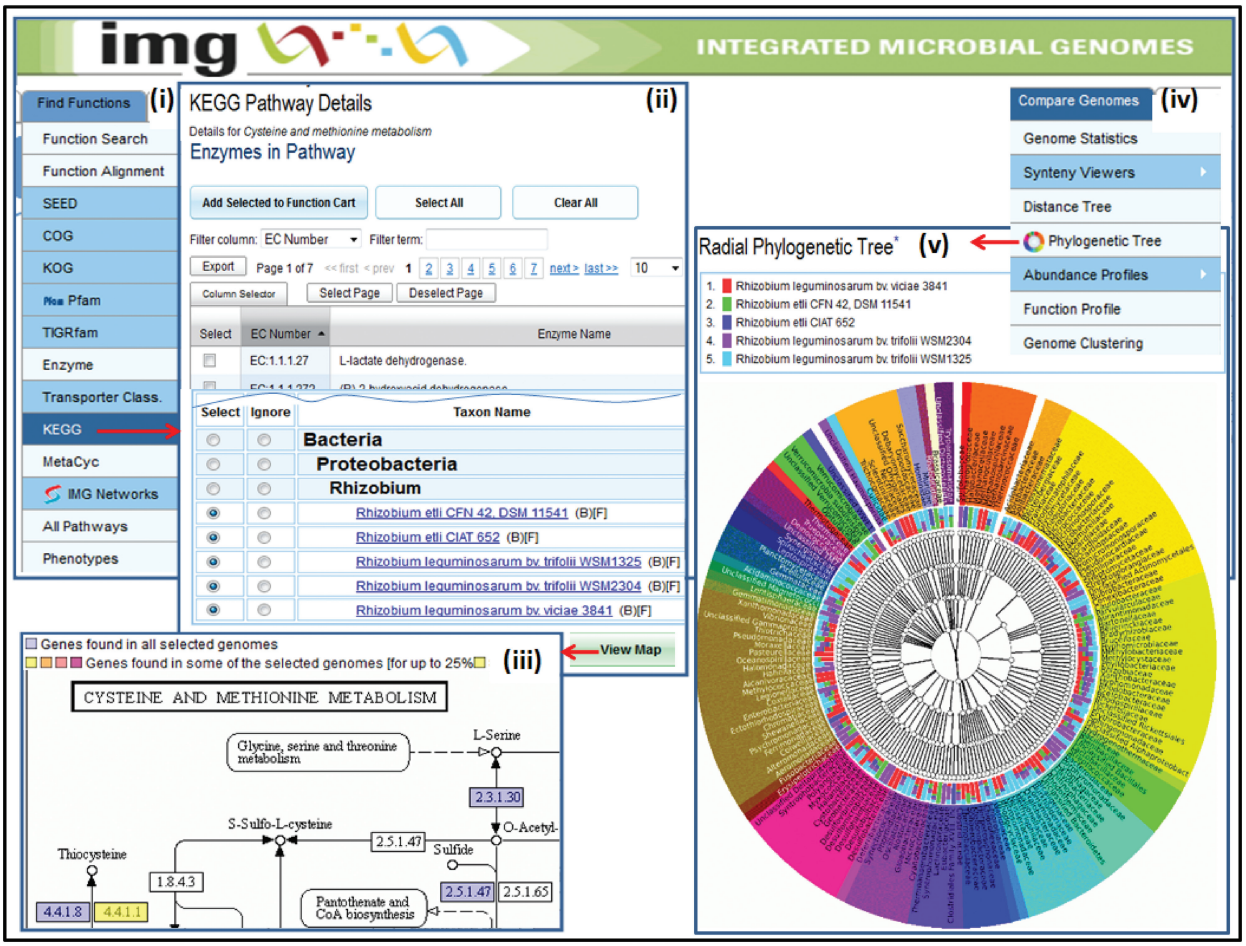


**Figure 4.** Comparative analysis tools. (**i**) A pathway is selected from the list of KEGG pathways via the KEGG option of the 'Find Functions' menu, and subsequently (**ii**) the 'KEGG Pathway Details' lists its associated enzymes and the list of genomes organized phylogenetically. (**iii**) Once genomes are selected for comparison, the result is displayed in the context of the KEGG pathway map, with each enzyme number on the map coloured depending on the percentage of genomes with a gene associated with that enzyme. (**iv**) The 'Radial Phylogenetic Tree' is one of several tools provided for comparing genomes, and (**v**) allows comparing the BLAST hits of the genes of up to five user selected genomes to the genes of all the genomes in the database using a colour-coded hierarchical circular tree viewer.

result displayed on the KEGG pathway map, as illustrated in Figure 4(iii). Each enzyme number on the map is coloured depending on the percentage of genomes with a gene associated with that enzyme, whereby the tooltip for a coloured enzyme displays the number of these genomes.

Genomes can be compared using two open source graphical viewers, 'Phylogenetic Distance Tree' and 'Radial Phylogenetic Tree', available under the 'Compare Genomes' main menu, as illustrated in Figure 4(iv). For both tools, genomes are selected for comparison from a list of genomes similar to that shown in Figure 4(ii). The 'Phylogenetic Distance Tree' computes the phylogenetic distance between genomes selected for comparison based on the 16S alignment derived from the SILVA database (29). For genes whose sequence is not included in the alignment the closest match is used, if the identify of it to the 16S gene of the IMG taxon is >97%. The distance tree is displayed using the Archaeopteryx tool (http://www.phylosoft.org/archaeopteryx/), which uses phyloXML for data exchange (30). Each node in the tree hyperlinked to the IMG genome page for that node.

The 'Radial Phylogenetic Tree' tool originally developed for MG-RAST (31), allows comparing the BLAST hits of the genes of up to 5 user selected genomes to the genes of all the genomes in the database using a colour-coded hierarchical circular tree viewer. This viewer displays the BLAST hits at different taxonomic levels, with more statistics for the hits for each genome provided by hovering the mouse over the nodes of the tree.

Genomes can be compared in terms of sequence conservation using VISTA tools (32), the Artemis comparison tool (33) and a 'Dotplot' tool which employs the program 'Mummer' to generate dotplot diagrams between two genomes.

In addition to the analysis tools available in IMG, IMG/ER provides tools for identifying and correcting annotation anomalies, such as dubious protein product names, and for filling annotation gaps detected using IMG's comparative analysis tools, such as genes that may have been missed by gene prediction tools or genes without predicted functions (24). Gene annotations that result from expert review and curation are captured in IMG/ER as so called 'MyIMG' annotations associated with individual scientist or group accounts, with curated genomes included into Genbank either as new submissions or as revisions of previously submitted data sets.

## FUTURE PLANS

IMG's genome sequence data content is maintained through regular updates from public sequence data resources. Since proteomics, transcriptomics, metabolomics and other 'omics' data are increasingly employed to refine our understanding of the functions of genes, additional types of 'omics' data will be gradually included into IMG following a similar integration approach and analysis tools to those developed for protein expression data.

IMG's integrated data framework allows assessing and improving the quality of genome annotations. Thus, the quality of gene models for genomes available in public resources is known to vary greatly depending on the quality of sequence and the software used for annotation. For example, an analysis conducted at JGI of the protein coding genes of microbial genes in Genbank indicates that ~10% (over 1 million) of predicted protein-coding are erroneous: they are false positive genes, unidentified pseudogene fragments or genes with translational exceptions, or have incorrectly predicted start sites. In order to improve the consistency of annotation and the quality of predicted genes, a project for the re-annotation of all public microbial genomes in IMG has been launched recently. This project relies on a gene quality assessment pipeline, GenePRIMP (34) that allows performing automated correction of gene models including insertion of missed genes, extension of 'short' genes and identification of putative pseudogenes.

The significant drop in the cost of sequencing has resulted in an exponential growth of new genome sequence data sets posing computational, data management and analytical challenges for the biological interpretation of these data sets. Furthermore, scientists are facing a data overload involving an increasing burden of analyzing a rapidly growing number of genomic data. These computational, data management and analytical challenges can be alleviated by synthesizing genomic data using the 'pangenome' conceptual abstractions (35). A pangenome consists of the core part of a species (i.e. the genes present in all of the sequenced strains or of all samples of a microbial community) and the variable part (the genes present in some but not all of the strains or samples). An experimental version of IMG has been extended with five pangenomes, as well as analysis tools and viewers that allow users to explore individual pangenomes and compare pangenomes and genomes. A public version of IMG containing pangenome data and analysis tools is expected to be released in the near future.

## ACKNOWLEDGEMENTS

We thank Henrik Nordberg, Roman Nikitin, Simon Minovitsky, Amrita Pati, Konstantinos Liolios and Ioanna Pagani for their contribution to the development and maintenance of IMG. The work of JGI's production, cloning, sequencing, assembly, finishing and annotation teams is an essential prerequisite for IMG. Eddy Rubin and James Bristow provided, support, advice and encouragement throughout this project.

## FUNDING

Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy (Contract No. DE-AC02-05CH11231); Office of Science of the U.S. Department of Energy (Contract No. DE-AC02-05CH11231, resources of the National Energy Research Scientific Computing Center) and US National Institutes of

Health Data Analysis and Coordination Center (Contract No. U01-HG004866, IMG-HMP system). Funding for open access charge: University of California.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
2. Liolios,K., Chen,I.M., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **38**, D346–D354.
3. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
4. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.*, **2**, 953–971.
5. Moller,S., Croning,M.D.R. and Apweiler,R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.
6. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
7. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
8. Griffiths-Jones,S., Moxon,S., Marshall,M., Khan-na,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
9. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
10. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
11. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunesekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
12. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
13. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daughterty,L., Duquenne,L. *et al.* (2005) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
14. Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
15. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, **38**, D355–D360.
16. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.
17. Ivanova,N.N., Anderson,I., Lykidis,A., Mavrommatis,K., Mikhailova,N., Chen,I.A., Szeto,E., Palaniappan,K., Markowitz,V.M. and Kyrpides,N.C. (2007) Metabolic reconstruction of microbial genomes and microbial community metagenomes. *Technical Report 62292*. Lawrence Berkeley National Laboratory. http://img.jgi.doe.gov/w/doc/imgterms.html (September 2011, date last accessed).
18. Mavromatis,K., Ivanova,N.N., Chen,I.A., Szeto,E., Markowitz,V.M. and Kyrpides,N.C. (2009) The DOE-JGI standard operating procedure for the annotations of microbial genomes. *SIGS*, **1**, 68–71, http://standardsingenomics.org/index.php/sigen/article/view/sigs632 (September 2011, date last accessed).
19. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
20. Saier,M.H. Jr, Yen,M.R., Noto,K., Tamang,D.G. and Elkan,C. (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res.*, **37**, D274–D278.
21. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *PNAS*, **96**, 2896–2901.
22. Mavromatis,K., Chu,K., Ivanova,N., Hooper,S.D., Markowitz,V.M. and Kyrpides,N.C. (2009) Gene context analysis in the integrated microbial genomes (IMG) data management system, accepted for publication. *PLoS ONE*, **4**, e7979, doi:10.1371/journal.pone.0007979.
23. Markowitz,V.M., Chen,I.A., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Anderson,I., Lykidis,A., Mavromatis,K. *et al.* (2009) The integrated microbial genomes (IMG) system: an expanding comparative analysis system. *Nucleic Acids Res.*, **38**, D382–D390.
24. Markowitz,V.M., Mavromatis,K., Ivanova,N.N., Chen,I.A., Chu,K. and Kyrpides,N.C. (2009) IMG ER: a system for microbial annotation expert review and curation. *Bioinformatics*, **25**, 2271–2278.
25. The Human Microbiome Jumpstart Reference Strains Consortium. (2010) A catalog of reference genomes from the human microbiome. *Science*, **328**, 994–999.
26. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, **462**, 1056–1060.
27. Unell,M., Abraham,P.E., Shah,M., Zhang,B., Rückert,C., VerBerkmoes,N.C. and Jansson,J.K. (2009) Impact of phenolic substrate and growth temperature on the *Arthrobacter chlorophenolicus* proteome. *J. Proteome Res.*, **8**, 1953–1964.
28. Florens,L., Carozza,M.J., Swanson,S.K., Fournier,M., Coleman,M.K., Workman,J.L. and Washburn,M.P. (2006) Analyzing chromatin remodelling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods*, **40**, 303–311.
29. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W., Peplies,J. and Glöckner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acid Res.*, **35**, 7188–7196.
30. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biologyand comparative genomics. *BMC Bioinformatics*, **10**, 356.
31. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The Metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
32. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
33. Carver,T.J., Rutherford,K.M., Berriman,M., Rajandream,M.A., Barrell,B.G. and Parkhill,J. (2005) ACT: the Artemis comparison tool. *Bioinformatics*, **21**, 3422–3423.
34. Pati,A., Ivanova,N.N., Mikhailova,N., Ovchinnikova,G., Hooper,S.D., Lykidis,A. and Kyrpides,N.C. (2010) GenePRIMP: A GENE PRediction IMprovement Pipeline for Prokaryotic genomes. *Nat. Methods*, **7**, 455–457.
35. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *PNAS*, **102**, 13950–13955.